



**HAL**  
open science

# Neural Signatures Of Musical And Linguistic Interactions During Natural Song Listening

Giorgia Cantisani, Shihab Shamma, Giovanni M Di Liberto

► **To cite this version:**

Giorgia Cantisani, Shihab Shamma, Giovanni M Di Liberto. Neural Signatures Of Musical And Linguistic Interactions During Natural Song Listening. 2024. hal-04529950

**HAL Id: hal-04529950**

**<https://hal.science/hal-04529950v1>**

Preprint submitted on 2 Apr 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---

# NEURAL SIGNATURES OF MUSICAL AND LINGUISTIC INTERACTIONS DURING NATURAL SONG LISTENING

---

**Giorgia Cantisani**

Laboratoire des systèmes perceptifs, CNRS  
École normale supérieure, PSL University  
Paris, France  
gcantisani@ens.fr

**Shihab A. Shamma**

Neural Systems Laboratory  
University of Maryland  
College Park, USA  
sas@umd.edu

**Giovanni M. Di Liberto**

ADAPT, Institute of Neuroscience  
Trinity College, The University of Dublin  
Dublin, Ireland  
diliberg@tcd.ie

## ABSTRACT

How are songs processed in the human brain? In song, tunes and lyrics are tightly bound in a music-language synergy to convey meaning and emotions beyond mere linguistic content, raising questions on how the two components are represented and integrated into a cohesive perceptual whole. Previous research pointed to areas of the human cortex sensitive to music, speech, and song, finding both shared and specialized sites. Yet, the interactions between tunes and lyrics processing when listening to songs remain poorly understood. To tackle this question, we probed neural predictive mechanisms specific to music and speech with electroencephalography. The encoding of melodic predictions was compared when listeners were presented with songs or the corresponding hummed (speech-free) melodies. Similarly, the encoding of phonemic predictions was studied in song and the corresponding spoken (melody-free) lyrics. We found that the concurrence of music and speech in songs alters how their predictive signals are generated and processed, altering their neural encoding. Furthermore, we found a trade-off in the neural encoding of melodic and phonemic expectations, with their balance depending both on who was listening (internal driver reflecting the listener's preference, e.g., musical training) and how the song is composed and performed (external driver reflecting the salience of lyrics and tunes). Altogether, our results indicate that song involves parallel prediction processes competitively interacting for the use of shared processing resources.

**Keywords** Song · Speech · Music · Predictive Processes · EEG

## Introduction

Music and speech are forms of communication central to human societies. Both music and speech are human universals that engage learning and cultural transmission. They make use of perceptually discrete elements embedded within highly structured temporal sequences in a principled way and can be produced with the earliest of musical instruments - the *human voice* [1]. Darwin was puzzled by the many links between music and speech and suggested that the evolutionary history of the two might be rooted in a protolanguage based on *song* [2]. Whether true or false, the development of musical and speech abilities has ancient origins, and neural pathways and mechanisms specific to music, much like speech, may thus have formed as part of human evolution [3, 4, 5, 6]. Yet, it remains uncertain if music and speech processing share a neural architecture that is comparable beyond their sensory and motor machinery or if the underlying systems share *general working mechanisms and fundamental principles*.

Puzzling is the case of singing, poetry, infant-directed speech, and all expression modalities where boundaries between speech and music are blurred. Tunes and lyrics are tightly bound and weighted in a music-language synergy (e.g., prosody in speech, melody in song), conveying emotions and intentions beyond the linguistic content and raising questions on how the two are integrated into a cohesive song percept. While such synergy points at lyrics-tunes integration mechanisms that extend beyond a shared sensorimotor system [7, 8], it is unclear how music and speech components are combined into a perceptual whole.

Here, we investigate if song perception emerges as the integration of two independent music and speech processes or if, and in what ways, the two processes interact with each other. In other words, does the simultaneous presence of music and speech properties in song alter how those two components are represented in the human brain? To answer that question, the present study probed music and speech encoding with electroencephalography (EEG), comparing the neural processing of song, music, and speech. Specifically, our investigation focuses on the study of music and speech expectation mechanisms, which have been found to reflect similar predictive processing principles and are taken here as a common ground for the investigation of the two processes.

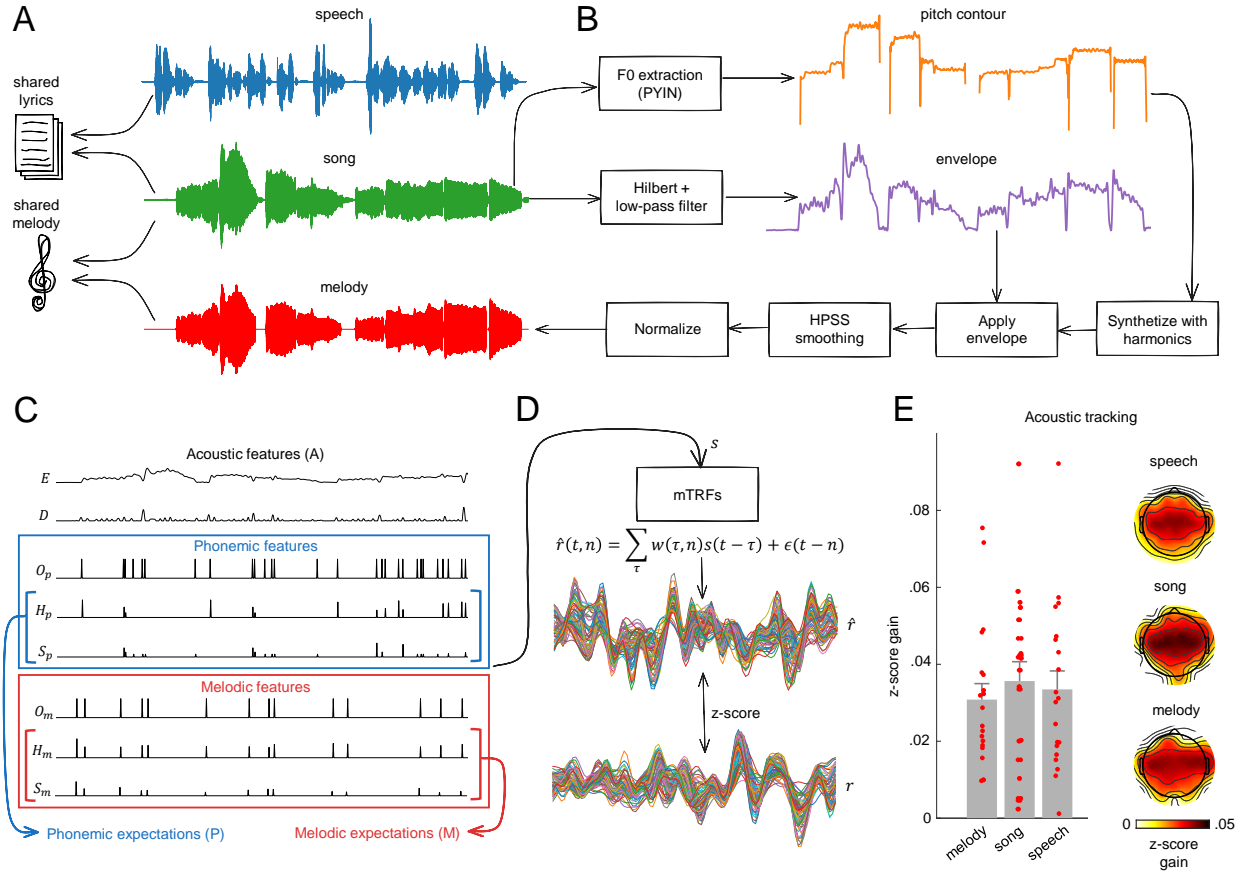
Previous comparative studies focused on identifying cortical areas responding to music and speech by using neuroimaging techniques with high spatial resolution, such as fMRI, finding both shared and specialized cortical sites [9, 7, 10, 8]. Interestingly, there is also evidence for non-primary cortical sites that are selective to song but not speech or music without lyrics [11]. Yet, while spatial overlap or segregation tells us about *where* the information is processed, it does not necessarily tell us *how* it is processed. For instance, neural activation in distinct cortical sites might be underpinned by a common predictive mechanism [12, 13, 14, 15, 16]. In fact, there is considerable evidence pointing to predictive processing as a foundational neural mechanism, which would underpin both music [17, 18, 19] and speech [20, 21, 22, 23, 24, 25, 26] perception, in spite of its manifestation occurring in spatially distinct cortical sites. Auditory prediction errors have been shown to modulate EEG signals, offering a window into top-down neural prediction mechanisms both for music [17, 18, 19] and speech [20, 21, 22, 23, 24, 25, 26]. Yet, the interaction between music and speech predictions in song processing remains unknown.

Here, we used electroencephalography (EEG) and linear encoding analyses to isolate neural signatures of music and speech prediction, testing if and how the two processes interact during song perception. To that end, we designed an experiment involving the listening of songs (sung speech with no instrumental music), melodies (hummed versions of the same songs with no speech content), and speech (spoken versions of the same songs without melodic and rhythmical content). Music-speech interactions were studied by contrasting the neural encoding of speech predictions during speech and song listening and the neural encoding of melodic predictions during music and song listening. Further analyses were conducted to determine if music-speech interactions depend on the listener’s musical training (internal factors) and on the song’s complexity (external factors). The results are in line with our hypothesis that speech and music predictions are carried out in parallel, with one prediction being influenced by the other and providing a new angle to the debate on the music-speech interaction in songs. Specifically, we found that their processing occurs in parallel while relying on common cognitive resources, leading to competitive neural dynamics that we could measure with EEG.

## Results

EEG data were acquired from  $N = 20$  adult participants listening to songs and the corresponding spoken and hummed versions. In doing so, the linguistic content - the *lyrics* - is shared between spoken and sung versions. Conversely, the melodic content - the *rhythm and pitch progressions* - is shared between hummed and sung versions (Fig. 1, A). Speech and song stimuli were actual recordings from the same singers, while melody ones were synthesized based on an ad-hoc song-to-hum algorithm (Fig.1, B).

Multivariate temporal response functions (mTRFs) were then estimated with multiple lag linear regression [27, 28] to assess the simultaneous cortical tracking of linguistic and melodic properties for the three types of stimuli separately (Fig. 1, C). Within this framework, brain responses can be described as a linear combination of multiple predictor variables (Fig. 1, D), thus enabling the neurophysiological interpretation of the relative model weights and giving insights into the neural encoding of particular stimulus dimensions [28, 29]. Here, we probed neural processes reflecting top-down predictive processes at the level of speech phonemes [30, 20, 26] and music notes [17, 18, 19]. mTRF models were fit to capture the neural encoding of stimulus properties at the level of sound acoustics ( $E$ : envelope,  $D$ : half-way rectified envelope derivative), segmentation ( $Op$ : phoneme onsets, and  $Om$ : note onsets), and statistical expectations ( $P$ : phonemic expectations;  $M$ : melodic expectations). Acoustic and segmentation features served as nuisance regressors, absorbing EEG variance relating to lower-level features and, in turn, magnifying our ability to assess the encoding of the higher-level top-down processes [24].



**Figure 1: Experimental, analytic framework, and envelope tracking validation.** (A) EEG signals were recorded from participants as they listened to 18 sung speech recordings (song), their natural speech counterparts (speech), and their hummed version (melody). Song shared the linguistic content with the speech stimuli and the melodic content with the melody stimuli. (B) Spoken and sung lyrics are real recordings, while melodies were synthesized based on an ad-hoc song-to-hum algorithm. (C) Sound acoustics, onsets, and expectation features were independently extracted from the three versions of the stimuli. (D) Multivariate lag regression was then used to estimate the linear relationship between different subgroups of such features and the EEG signal. The resulting multivariate temporal response functions (mTRFs) were then studied in terms of how well they could predict unseen EEG data measured as z-scored Pearson’s correlation and by observing which features and peri-stimulus latencies were driving such predictions. (E) Bar plots report the z-score gain (average across channels) over the baseline for each stimulus type, averaged across all subjects (mean  $\pm$  SE) when considering an mTRF using only acoustic features (envelope and its derivative) as regressors. Red dots represent individual participants. Z-score gains were significant for all conditions (Bonferroni-corrected with  $m = 2$  Wilcoxon rank-sum test for melody and  $t$ -test for song and speech, melody:  $p < 0.001$ ,  $Z = 3.92$ ; song:  $p < 0.001$ ,  $d = 1.63$ ; speech:  $p < 0.001$ ,  $d = 1.60$ ), while there was no statistical evidence for differences across conditions (one-way ANOVA,  $p > 0.05$ ). Topoplots indicate z-scores across the scalp for the average participant. Only significant channels (FDR corrected  $t$ -test with significance threshold set at  $p < 0.001$ ) are displayed.

Melodic surprisal  $S_m$  and entropy  $H_m$  (together referred to as melodic expectation features  $M$ ) were calculated with IDyOM - Information Dynamics of Music - [31], a generative model of musical structure that causally estimates the probability distribution over all possible note-pitch continuations given short-term information about the melodic sequence and long-term information from a large corpus of Western music. The first estimate is associated with the short-term memory of the statistics of the musical piece under test, while the second simulates the statistical knowledge of a listener implicitly acquired in a lifetime of exposure to music. Based on such estimates, we could define two measures that capture distinct aspects of expectation and surprise at each new note within a melody: note-pitch entropy and surprisal. These two features are complementary in that the first indicates the degree of expectation (*after* the event is observed), while the second clarifies the degree of uncertainty of the context (*before* the event is observed), thus corresponding to a weaker or stronger sequence violation, respectively.

On the other hand, phonemic surprisal  $S_p$  and entropy  $H_p$  (together referred to as phonemic expectation features  $P$ ) were derived from probability distributions learned from an English pronouncing dictionary [32]. Note that expectations for the phonemic and melodic streams were consistent in two ways. First, their calculation is in line with the information theory definitions of Shannon entropy and information content [33]. Second, melodic expectations only consider the pitch value associated with each note, excluding considerations of onset time expectations [17]. This analytic decision allowed us to study melodic expectations that conceptually align with speech features, in that phonemes' expectations were also built disregarding timing information.

### Cortical tracking of speech, song, and melody acoustics

First, mTRF analyses were carried out to determine the cortical tracking of speech, melody, and song acoustic features, providing a baseline for evaluating differences between conditions. Speech and music have strong acoustic differences, and depending on the set of stimuli, one could measure stronger or weaker envelope tracking simply due to different signal-to-noise ratios [34]. Moreover, lower-level differences in the acoustics of music and speech could falsely induce differences in response strengths and patterns of higher-level linguistic and melodic components. To avoid such confounds, we constructed our stimuli to minimize such spectral differences and other substantial acoustic mismatches due, for instance, to the use of musical instruments other than the human voice.

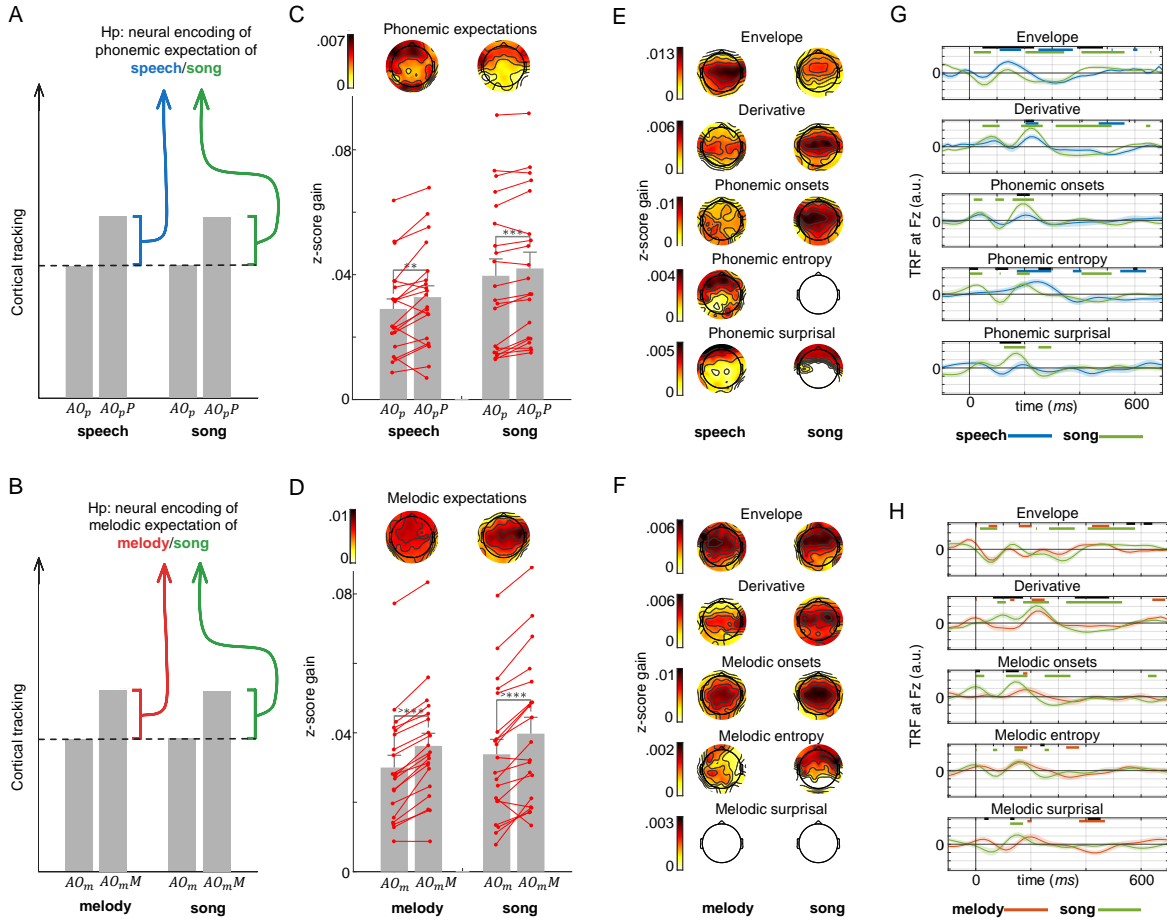
We observed a robust low-frequency ( $1-8\text{Hz}$ ) cortical tracking of the sound acoustics ( $A$ : envelope  $E$  and its derivative  $D$ ) for the three types of stimuli. EEG prediction correlations were above chance (Bonferroni-corrected Wilcoxon rank-sum test for melody and  $t$ -test for speech and song, melody:  $p < 0.001$ ,  $Z = 3.92$ ; song:  $p < 0.001$ ,  $d = 1.63$ ; speech:  $p < 0.001$ ,  $d = 1.60$ ; Fig. 1, E), with the chance level calculated by running the mTRF analysis after randomly shuffling the input (*i.e.*, stimulus and EEG were misaligned). In all the conditions, best-predicted electrodes were in a broad centro-parietal area of the scalp (FDR corrected  $t$ -test with significance threshold set at  $p = 0.001$ ; Fig. 1, E). There was no statistical evidence for differences across conditions on the average of electrodes (one-way ANOVA,  $p > 0.05$ ; Fig. 1, E) and also on individual electrodes (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ).

### Cortical tracking of phonemic expectations during speech and song listening

EEG responses to speech and song were used to assess the cortical encoding of phonemic expectations and test the hypothesis that those are encoded in responses to songs as they do for speech (Fig. 2, A). For both speech and song conditions, including  $P$  among the predictors significantly improved the average model predictions across channels ( $t$ -test, speech:  $p = 0.005$ ,  $d = 0.73$ ; song:  $p < 0.001$ ,  $d = 0.92$ ; Fig. 2, C) but also almost all individual channel predictions (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ; Fig. 2, E). Significance was assessed by comparing the predictive power of the full model  $AOpP$  to the one of a reduced model  $AOp$  representing the null hypothesis, in which the predictor under investigation  $P$  was randomly shuffled, thus controlling for correlated variables, namely acoustics  $A$  and phoneme onsets  $Op$ . Thanks to this analysis, we could determine that phonemic expectations explained a distinct portion of the data variance, thus accounting for unique features of brain responses.

We also tested the individual contribution of phonemic entropy  $H_p$  and surprise  $S_p$  to disentangle their effects as we hypothesize the two variables reflect separable and yet complementary stages of information processing. When testing for surprisal and entropy features individually (*i.e.*, shuffling only one of the two at the time), we found that, when considering the average over all electrodes, both were encoded in responses to song ( $t$ -test,  $S_p$ :  $p = 0.049$ ,  $d = 0.48$ ;  $H_p$ :  $p = 0.039$ ,  $d = 0.51$ ) and speech ( $t$ -test,  $S_p$ :  $p = 0.024$ ,  $d = 0.56$ ;  $H_p$ :  $p = 0.007$ ,  $d = 0.69$ ) with no significant difference between the encoding of the two ( $t$ -test,  $p > 0.05$ ). However, when looking at individual channels, we had significant z-score gains for entropy and surprisal individually only for speech (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ; Fig. 2, E). For song, only considering the two complementary features together led to the significant reconstruction of individual channels (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ; Fig. 2, C), suggesting an interaction between the two. Finally, especially for the speech condition, phoneme-related features seem to better reconstruct left channels (Fig. 2, panels C and E) as opposed to simple acoustics (panel E of Fig. 1 and 2). Yet, we found no significant effect of left-lateralization of phoneme-related features across subjects ( $p > 0.05$ ; lateralization test).

Note that the encoding of the control variables ( $AOp$  model) on the average over all electrodes was already significant per se ( $t$ -test or Wilcoxon signed-rank test depending on Gaussianity;  $E$ : speech:  $p < 0.001$ ,  $d = 1.12$ ; song:  $p = 0.001$ ,  $d = 0.87$ ;  $D$ : speech:  $p = 0.012$ ,  $Z = 2.5$ ; song:  $p < 0.001$ ,  $Z = 3.92$ ;  $Op$ : speech:  $p = 0.001$ ,  $d = 0.89$ , song:  $p < 0.001$ ,  $Z = 3.92$ ). Interestingly, for speech, we found a better fit due to  $E$  than  $D$ , indicating a stronger encoding of the energy envelope as opposed to the acoustic onsets ( $t$ -test,  $p < 0.05$ ,  $d = 0.65$ ). Conversely, for song, we found a better fit due to  $D$  than  $Op$ , indicating a stronger encoding of categorical onsets as opposed to acoustic ones ( $t$ -test,  $p < 0.01$ ,  $d = 0.69$ ). This result is also reflected in the comparison across the song and speech conditions: for speech,



**Figure 2: Cortical tracking of phonemic and melodic expectations during song listening.** (A) Hypothesis 1: we can robustly track phonemic expectations in the EEG responses to song as we do for speech-only. (B) Hypothesis 2: we can robustly track melodic expectations in the EEG responses to song as we do for melody-only. (C-D) Z-transformed encoding correlations for individual channels (scalp topographies, FDR corrected  $t$ -test with significance threshold set at  $p < 0.05$ ) and for the average over all channels (bar plot, mean+SE across participants and trials) for pairs of conditions and two subsets of features, A (acoustic + expectation features shuffled) and AP (acoustic + expectation features). Red dots represent individual participants. For both phonemic and melodic expectations, we had significant z-score improvement for all conditions when adding the expectation features (surprise and entropy). For phonemes,  $p = 0.005$  with  $d = 0.73$  for speech,  $p < 0.001$  with  $d = 0.92$  for song. For music notes,  $p < 0.001$  with  $d = 2.22$  for melody,  $p < 0.001$  with  $d = 1.16$  for song. (E-F) EEG prediction correlations gain with respect to the null model displayed on the scalp for pairs of conditions and different features. Only significant channels (FDR corrected  $t$ -test with significance threshold set at  $p < 0.05$ ) are displayed. (G-H) TRF kernels relative to the envelope, its derivative, onsets, entropy, and surprisal at Fz (mean+SE across participants) for pairs of conditions. Horizontal bars indicate the significance of TRF's portions ( $t$ -test with significance threshold set at  $p < 0.05$ ) for each condition (colored bars) and their difference (black bars).

we found a stronger envelope encoding for right temporal electrodes, and, for song, a stronger onset encoding for centro-frontal electrodes (FDR-corrected  $t$ -test with significance threshold set at  $p = 0.05$ ).

Finally, panel G of figure 2 juxtaposes TRFs at channels Fz for pairs of conditions and features of interest to assess if the same feature is processed with different timing, polarities, and strengths for speech and sung speech. Moreover, different features may exhibit different dynamics that reflect different stages of linguistic processing. Entropy relates to the uncertainty of the prediction of the upcoming event before observing it. Thus, it describes the overall probability distribution over all possible continuations. Surprise is extracted from that same distribution but is specific to the observed event. Therefore, it is likely that this dynamic will be reflected in the respective TRFs. Similarly, acoustic processing should anticipate any higher linguistic processing. The temporal order we observe is indeed consistent with these considerations, with surprisal to exhibit an earlier main peak than entropy (for speech: 170ms vs 200ms and for song: 190ms vs 200ms). The entropy peaks are at the same latency for both speech and song and are time-locked with

categorical onsets as opposed to surprisal, which anticipated categorical onsets. In general, it is evident that delays and polarities for acoustic features and categorical onsets do not change between speech-only and sung speech, while they do change when it comes to phonemic expectation features.

### Cortical tracking of melodic expectations during humming and song listening

Analogous to the analysis of speech expectations above, EEG responses to melodies and songs were used to assess the cortical encoding of melodic expectations and test the hypothesis that those are encoded in the responses to songs as they are for melodies (Fig. 2, B). For both melody and song conditions, including  $M$  among the predictors significantly improved the average model predictions across channels ( $t$ -test, melody:  $p < 0.001$ ,  $d = 2.22$ ; song:  $p < 0.001$ ,  $d = 1.16$ , Fig. 2, D), but also almost all individual channel predictions (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ; Fig. 2, D), meaning melodic expectations explain a distinct portion of the variance in the data. Significance was assessed by comparing the predictive power of the full model  $AOmM$  to that of a model representing the null hypothesis  $AOm$ , in which the predictor under investigation  $M$  was randomly shuffled, thus controlling for correlated variables.

When testing for the individual contribution of melodic entropy  $Hm$  and surprise  $Sm$ , we found that only entropy was significantly encoded in responses to melodies and songs, both considering the average across electrodes ( $t$ -test, melody:  $p = 0.046$ ,  $d = 0.49$ ; song:  $p < 0.001$ ,  $d = 0.95$ ) and individual channels (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ; Fig. 2, F). For both melody and song conditions, mostly centro-frontal channels were significantly reconstructed with  $Hm$  but none with  $Sm$ . However, when combining the two features, all channels became significant (Fig. 2, D). Moreover, looking at fit improvement, the z-score gain obtained with the combined set of expectation features is not the sum of the gains obtained with entropy and surprisal individually, suggesting an interaction between the two variables, which might be more than simply additive.

Interestingly, the encoding of melodic expectations is significantly right-lateralized only in responses to songs ( $p = 0.038$ ; lateralization test; Fig. 2, D). A potential confound might be that EEG data are biased towards one hemisphere. However, the lateralization test on lower-level features ( $E$ ,  $D$ , and  $Om$ ) indicated no significant difference in the predictive power in the left and right hemispheres ( $p > 0.05$ ; lateralization test). Thus, together with the result that significant lateralization was observed only for brain responses to song, it suggests that the bias toward one hemisphere was minimal and that the lateralization effect is due to the simultaneous processing of lyrics and tunes.

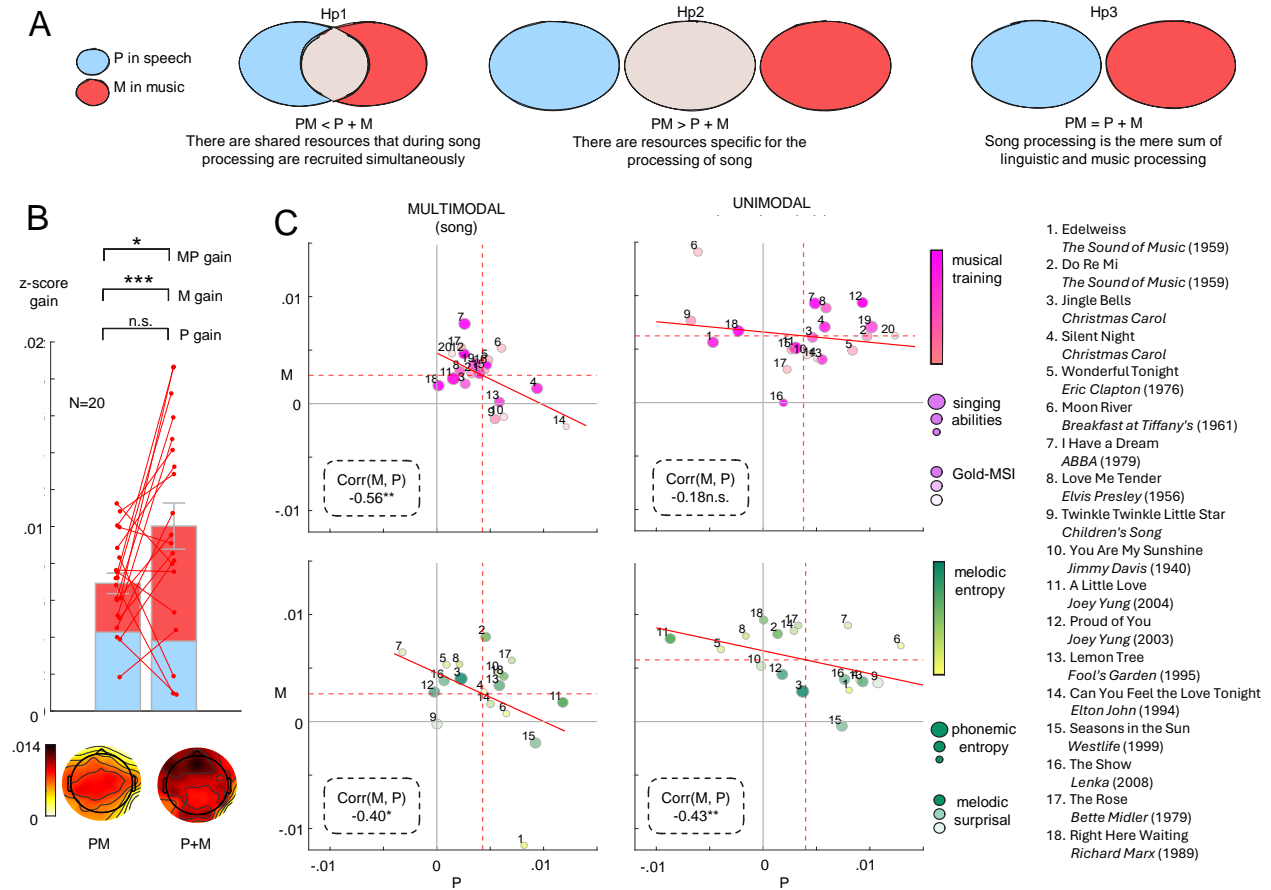
Note that the encoding of the control variables was already significant per se on the average over all electrodes ( $AOm$  model,  $t$ -test or Wilcoxon signed-rank test depending on Gaussianity,  $E$ : melody  $p < 0.001$ ,  $Z = 3.47$ , song  $p = 0.006$ ,  $d = 0.72$ ;  $D$ : melody  $p = 0.003$ ,  $Z = 2.95$ , song  $p < 0.001$ ,  $Z = 3.92$ ;  $Om$ : melody  $p < 0.001$ ,  $d = 1.36$ , song  $p < 0.001$ ,  $d = 1.55$ ). Additionally, we found a higher fit improvement due to  $D$  than  $Om$ , indicating a stronger encoding of categorical onsets as opposed to acoustic ones for both melody and song ( $t$ -test, melody:  $p = 0.049$ ,  $d = 0.45$ ; song:  $p < 0.039$ ,  $d = 0.51$ ), and a higher fit improvement due to  $E$  than  $D$  indicating a stronger encoding of the envelope than to its derivative for melody ( $t$ -test,  $p = 0.007$ ,  $d = 0.69$ ).

We found no significant differences across the song and melody conditions for the encoding of individual features except for electrodes in the right temporal area for entropy and centro-frontal electrodes for acoustic onsets. In both cases, the encoding is stronger in the song condition (FDR corrected  $t$ -test with significance threshold set at  $p < 0.05$ ).

Finally, panel H of Fig. 2 displays TRFs at Fz for pairs of conditions and features of interest. Again, we hypothesize that different features do not present the same dynamics as they reflect different stages of music processing, ranging from low-level acoustic features to melodic structures. Therefore, expectations are presumed to continue to build up when listening to longer contexts. As for phonemic expectation, we anticipate that surprisal's and entropy's TRFs to reflect different stages of the statistical processing of a given note, inducing potentially different latencies in their respective TRFs. The temporal order we observe is consistent with these considerations: envelope TRFs exhibit a first negative deflection N1 at 60ms followed by a positive peak P1 at 125ms consistent with previous results obtained with a reduced model with only the envelope [34]. Acoustic responses to both melodies and songs show a consistent peak at a later latency of around 230ms while categorical onsets TRFs start differentiating more between the two conditions (song shows a P1 near 140ms while melody at 190ms).

### Phonemic and melodic expectations in responses to song and their interactions

Finally, EEG responses to song were used to assess the cortical encoding of melodic and phonemic expectations when they have to be generated in parallel (denoted as  $PM$ ) and to compare their total contribution with the superposition of the unimodal cases (denoted as  $P + M$ ), *i.e.*, of the phonemic expectations in responses to speech ( $P$ ) and the melodic expectations in responses to melodies ( $M$ ). This analysis and its terminology are inspired by studies of multisensory



**Figure 3: Phonemic and melodic expectations in responses to song and their interactions** (A) hypotheses on the interaction of the speech and music processing modules: If  $PM < P + M$ , there are shared neural resources that, during song processing, are recruited simultaneously (hypothesis 1). If  $PM > P + M$ , there are neural resources specific for the processing of songs (hypothesis 2). If  $PM = P + M$ , song processing is the mere sum of linguistic and music processing (hypothesis 3). (B) The z-scored fit improvement due to linguistic and melodic expectation in song ( $PM$ ) is compared to the summed fit improvement in the unimodal stimuli ( $P + M$ ).  $P + M$  is significantly larger than  $PM$  ( $t$ -test,  $p < 0.024$ ,  $d = 0.56$ ), and so is  $M$  within  $P + M$  ( $t$ -test,  $p < 0.001$ ,  $d = 1.17$ ). Note that both  $PM$  and  $P + M$  fit improvement were significant ( $t$ -test,  $PM$ :  $p < 0.001$ ,  $d = 2.85$ ,  $P + M$ :  $p < 0.001$ ,  $d = 1.84$ ) and so their components  $P$  ( $t$ -test, in  $PM$ :  $p < 0.001$ ,  $d = 1.56$ , in  $P + M$ :  $p = 0.005$ ,  $d = 0.73$ ) and  $M$  ( $t$ -test, in  $PM$ :  $p < 0.001$ ,  $d = 1.11$ , in  $P + M$ :  $p < 0.001$ ,  $d = 2.22$ ). Below the barplot are reported the fit improvement due to  $PM$  and  $P + M$  for individual channels (only significant channels are displayed, FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ). (C) the relative contribution of  $M$  plotted against the one  $P$  in the multimodal (song) and unimodal (speech and melody) conditions. In the first row, data points represent individual subjects (average performance over trials and channels), while in the second row, data points represent individual songs (average performance over subjects and channels). Data points color, size, and transparency reflect behavioral or stimulus measures. The correlation between  $M$  and  $P$  is reported for each scatterplot, with significance computed with a jackknife cross-validation.

integration [35]. Unimodal here refers to the condition where only linguistic or melodic information is present in the stimulus, such as in speech and melodies; multimodal refers to the condition when both are present, such as in songs. Consequently, three different hypotheses are tested regarding the nature of interactions between the speech and music processing modules in song perception (Fig. 3, A):

- If  $PM < P + M$ , there are shared neural resources between the speech and music processing modules that, during song processing, are simultaneously recruited.
- If  $PM > P + M$ , there are neural resources specific to the processing of songs.
- If  $PM = P + M$ , song processing is the mere sum of linguistic and music processing.

To assess these alternatives, we consider the contribution of  $P$  and  $M$  to the reconstruction of EEG song responses in a model including all features ( $AOpOmPM$ ), compared to the sum of the separate contributions from  $P$  or  $M$  to the



reconstruction of EEG responses to *speech* and *melodies*, with the unimodal linguistic ( $AO_{pP}$ ) and melodic ( $AO_{mM}$ ) models. The results evidently support the first hypothesis, with  $PM < P + M$  ( $t$ -test,  $p = 0.024$ ,  $d = 0.56$ , Fig. 3, B). This difference is driven by frontal channels (FDR corrected  $t$ -test with significance threshold set at  $p = 0.05$ ) and by the contribution of  $M$ , which is stronger in responses to melodies than to songs ( $t$ -test,  $p < 0.001$ ,  $d = 1.17$ ). For  $P$ , there is no significant difference between the multimodal and unimodal conditions ( $t$ -test,  $p > 0.05$ ). Note that both  $PM$  and  $P + M$  fit improvements were significant ( $t$ -test,  $PM$ :  $p < 0.001$ ,  $d = 2.85$ ,  $P + M$ :  $p < 0.001$ ,  $d = 1.84$ ) and so were their components  $P$  ( $t$ -test, in  $PM$ :  $p < 0.001$ ,  $d = 1.56$ , in  $P + M$ :  $p = 0.005$ ,  $d = 0.73$ ) and  $M$  ( $t$ -test, in  $PM$ :  $p < 0.001$ ,  $d = 1.11$ , in  $P + M$ :  $p < 0.001$ ,  $d = 2.22$ ).

### Internal and external drivers of shared resource allocation: subject preference and lyrics/tunes salience

The fact that fit improvement due to the combined linguistic and melodic expectations ( $PM$ ) differs from the summed fit improvement in the unimodal stimuli ( $P + M$ ) suggests a nonlinear lyrics-tunes integration, at least at the phonemic and melodic level, with shared resources between the speech and music processing modules that are allocated based on some internal strategy and intrinsic properties of the stimuli, i.e., a balance of salience between lyrics and tunes induced by a variety of features ranging from the acoustic to the semantic. This balance is assumed to be formulated by the composer during song creation or by the singer during performance. Its value is in making either the linguistic or melodic information more attractive and noticeable. We shall refer to this effect as a *bottom-up external driver/bias*. By contrast, an internal strategy refers to the listener’s *top-down internal driver/bias* - a preference for the melodic or linguistic content due to her musical interests, culture, training, mood, and other complex factors that affect the melodic and phonemic encoding. Here, we tested if such biases existed while being agnostic about their causes by correlating the relative contributions of  $P$  and  $M$  in the unimodal and multimodal conditions. The hypothesis is that  $P$  and  $M$  act jointly to enhance the use of the shared resources based either on subject-specific internal factors or/and song-specific external factors.

Figure 3C illustrates the relative contribution of  $M$  versus  $P$  in the multimodal (song) and unimodal (speech or melody) conditions. In the first row of scatterplots, each data point represents an individual subject (performance averaged over trials and channels); in the second row, data points represent individual songs (performance averaged over subjects and channels). For the subject-wise analysis,  $M$  and  $P$  relative contributions are *negatively* correlated only in the multimodal condition (Pearson’s correlation and jackknife cross-validation,  $r = -0.56$ ,  $p = 0.006$ ), while they are not in the unimodal condition ( $r = -0.18$ ,  $p > 0.05$ ). Thus, subjects exhibit an internal preference for melodic or linguistic content only in responses to songs. This is consistent with the hypothesis of shared resources that are allocated based on a strategy when melodic and linguistic information has to be processed simultaneously. By contrast, for individual songs, we observe a negative correlation in both the multimodal ( $r = -0.40$ ,  $p = 0.004$ ) and the unimodal case ( $r = -0.43$ ,  $p = 0.003$ ). The fact that the correlation does not drop in the unimodal condition suggests that lyrics and tunes competitively attract attention to one another because of how they are built and the intrinsic properties they also exhibit when the exact linguistic or melodic content is heard in isolation, as in the unimodal case.

We collected behavioral data from the participants through a questionnaire, which allowed us to test whether subjective preferences for either tunes or lyrics could be driven by factors such as musical training and perceptual abilities or by more general musical sophistication. We used the Goldsmiths Musical Sophistication Index (Gold-MSI), which can be factored into various facets of music sophistication: Active Engagement (AE), Perceptual Abilities (PA), Musical Training (MT), Singing Abilities (SA), and Emotions (EM) factors [36]. When considering centro-frontal electrodes, all indices negatively correlated with  $P - M$  in the multimodal case (AE:  $r = -0.41$ ,  $p > 0.05$ ; PA:  $r = -0.58$ ,  $p = 0.01$ ; MT:  $r = -0.49$ ,  $p = 0.011$ ; SA:  $r = -0.49$ ,  $p = 0.035$ ; EM:  $r = -0.51$ ,  $p = 0.001$ ; Gold-MSI:  $r = -0.50$ ,  $p = 0.023$ ) while in the unimodal case correlation scores are close to zero. When correlating the indices with  $M$ , correlation scores are positive in the multimodal case but only significant for the EM ( $r = -0.52$ ,  $p = 0.01$ ) and near zero in the unimodal case. Conversely, such indices are negatively correlated with  $P$  in the multimodal (AE:  $r = -0.45$ ,  $p = 0.025$ ; PA:  $r = -0.55$ ,  $p = 0.013$ ; MT:  $r = -0.56$ ,  $p < 0.001$ ; SA:  $r = -0.58$ ,  $p = 0.002$ ; EM:  $r = -0.56$ ,  $p = 0.002$ ; Gold-MSI:  $r = -0.56$ ,  $p = 0.002$ ) and close to zero in the unimodal one. Thus, a weaker phonemic expectation encoding in song responses corresponds to a higher music sophistication. Note, however, that all indices are correlated, making it hard to disentangle each factor’s contribution.

Finally, regarding the external drivers affecting the encoding of melodic and phonemic features, we hypothesized that statistical properties of the lyrics and tunes might affect their balance, i.e., the balance between the strength of the encoding of  $P$  and  $M$ . To do so, we tested if the difference between the encoding of those two properties ( $P - M$ ) correlated with the average melodic or phonemic entropy and surprise of a song. This analysis was motivated by previous studies showing that entropy and surprise predict musical pleasure [37], reward [38], and complexity [39]. Interestingly, complexity has been shown to be related to music enjoyment with an inverted U-shaped relationship [40, 13]. We found a significant correlation between ( $P - M$ ) and the average melodic surprisal only in the unimodal

case for centro-frontal electrodes ( $(P - M)$ :  $r = -0.59$ ,  $p = 0.002$ ;  $P$ :  $r = -0.59$ ,  $p = 0.001$ ;  $M$ :  $r = 0.39$ ,  $p > 0.05$ ). We then analyzed the songs in the dataset (no brain activity is involved here) to check if there is any relationship between the average entropy and surprisal. Such analysis reveals that songs having low phonemic surprisal also have low phonemic entropy ( $r = 0.98$ ,  $p < 0.001$ ), while in melodies, entropy and surprisal are less inter-dependent ( $r = 0.31$ ,  $p > 0.05$ ). In addition, melodic entropy is highly correlated with phonemic entropy ( $r = 0.95$ ,  $p < 0.001$ ), while melodic surprisal is not correlated with phonemic surprisal ( $r = 0.29$ ,  $p > 0.05$  respectively). This suggests that melodic surprisal is an expressive tool composers use to direct attention to specific elements of a composition and that entropy at the phonemic and melodic levels is apparently balanced. Yet, an extensive analysis of a large music corpus is necessary to assert if such correlations generalize over the stimuli used in this study.

## Discussion

In this study, we provide the first direct comparison of music and speech predictive mechanisms in a natural listening task. We probed melodic and phonemic predictions in EEG responses of participants listening to songs as well as the corresponding hummed melodies and spoken lyrics. Our results demonstrate a *robust*, *simultaneous*, and *distinct* cortical encoding of both melodic and phonemic predictions during song listening. This result goes beyond previous evidence obtained separately for speech and music [17, 18, 19, 20, 23, 24, 25, 22, 21] by considering the case where the two jointly occur. Crucially, we investigated the simultaneous unfolding of the two processes, discovering interactions of a competing nature that point to a shared cortical infrastructure. Based on those findings, we propose a high-level music-speech integration model, accounting for expression modalities at their intersection, like poems and songs.

### Comparing music and speech predictive mechanisms and their interactions with naturalistic songs

The role of prediction in speech and music listening differs radically. In speech, building linguistic expectations, such as phonemic predictions, can impact understanding (e.g., phonemic restoration, semantic interpretation) [41, 42, 43], especially in noisy, multi-talker environments [44, 45]. In music, expectations instead have been linked to emotions and engagement [17, 46, 47, 38, 37, 48]. Therefore, it remains unclear to what extent predictions in the two domains are related and comparable. Neural signatures of expectations observed in previous studies might reflect domain-specific computations based on the same statistical learning principles, but they could also involve domain-general mechanisms such as working memory or cognitive control [8, 49]. An additional complication is that predictions are generated at multiple time scales, reflecting the different structural levels of speech and music. Hence, one needs to select stimulus properties providing the most direct comparison across these domains [50, 15] while avoiding confounding factors like acoustic differences [34, 51, 52, 53]. All these factors highlight the need for ad-hoc comparative studies where speech and music are contrasted systematically across distinct abstraction levels (e.g., acoustics vs. phonemic units) while controlling for acoustic mismatch [51].

Neurophysiological research on sensory predictions largely relied on carefully crafted stimuli, where unexpected events are prominently and artificially placed [54, 55, 56, 57, 58]. Recent studies have indicated that neural prediction mechanisms can also be probed in response to the more subtle changes in expectations that are naturally part of speech [20, 21, 22, 23, 24, 25] and melodies [17, 18, 19], employing tasks that involve continuous listening without artificial manipulations. This approach enables measurements of multiple neural correlates simultaneously [27, 28] with continuous ecologically valid stimuli, which is advantageous compared to the more controlled paradigms where isolating a specific property (e.g., phonemes or notes) requires a dedicated experiment. The song-listening task here was designed to engage listeners in an ecologically valid manner, making our findings on the cortical encoding of the music-speech interplay relevant to our daily experiences.

### Disentangling influences of low- from high-level features

Our results show that the encoding of acoustic properties is distinct from that of phonemes and musical notes and their expectations, pointing to three levels of information processing, from acoustic to discrete units and corresponding prediction. The processing of linguistic and music units and their expectations is supported by numerous electrophysiological studies [17, 18, 19, 20, 59, 23, 24, 25, 22, 21], with the challenge being to untangle these responses from those to acoustics [60].

Results both on speech and music processing suggest that the relative strength of the encoding of low- vs. higher-order properties changes when listening to songs when both speech and music are present. In speech, the acoustic envelope is more prominently encoded than phonemic onsets, while the opposite is true in songs. In music, we found a stronger encoding of note onsets as opposed to the acoustic envelope for both melody and song. These differences suggest that the brain re-adjusts linguistic processing depending on the presence of music and, vice versa, adapts how music is

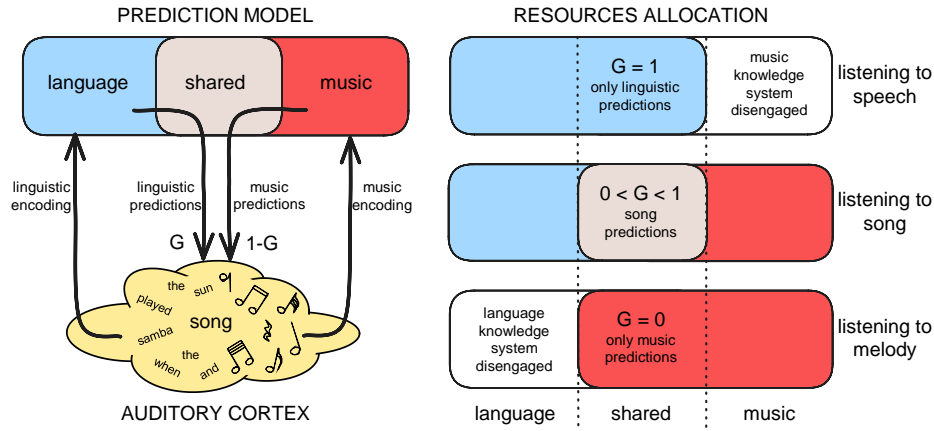


Figure 4: A **predictive processing model of speech-music integration**. The model entails two modules, one for language and one for music, which exhibit both specialized and shared processing. When language and music have to be processed simultaneously, like in song listening, linguistic and melodic predictions are carried out in parallel but are conditioned on each other, generating *song* predictions of the lyrics and tunes. This model is valid for all conditions we tested: music-only, speech-only, and song. In the unimodal case, either the right or left side of the model is inactivated. In song, the music and speech processing modules interact competitively for those shared resources, which are then allocated based on a gain  $G$  determined by complex internal and external factors.

processed when it needs to process speech information. However, we need to keep in mind that while for percussive instruments like a piano, acoustic onsets are good representatives of note transitions, for singing voice (intended as an instrument, with lyrics or hummed), energy fluctuations are mostly used as an expressive tool. Indeed, energy bursts are present at the start of phrases or at *intended* loudness changes. Thus, in song, envelope and onsets represent two distinct properties, one linked to melody unfolding and one to prosody.

### Simultaneity of speech and music in song reveals a competition between linguistic and melodic predictions

Our results indicate that the concurrence of music and speech in song listening alters the corresponding neural predictive processes compared to the unimodal scenarios, thus revealing their interactions. Overall, we found that the sum of the cortical encoding of the unimodal stimuli is larger than the encoding of the multimodal song ( $P + M > PM$ ), suggesting that the two processes share neural resources rather than relying on distinct neural infrastructures. Furthermore, our evidence points to both spatial and temporal changes, revealing a re-arrangement of how speech and music are processed when co-occurring in song. These results are consistent with the hypothesis that parallel speech and music predictions in song influence each other.

This also results support previous evidence on the existence of both specialized and shared neuronal populations for speech and music processing [7, 8, 10, 11, 9]. In song listening, the shared population would either have to multiplex between the two processes, favoring one at a time, or to process emerging song-specific features [11]. The latter observation is supported by the inverse relationship between speech and music encoding in song (Fig. 3). Song processing therefore is not merely "tunes + lyrics" processing [61, 62, 63, 52].

### Modelling the competitive nature of speech and music predictions

We propose a model that integrates music and speech predictive processes via two modules (Fig. 4), one for language and another for music, with both specialized and shared processes. The shared processes likely rely on a common neural infrastructure, for instance, song-selective neural population [11]. When both language and music are present, linguistic and melodic predictions are carried out in parallel but condition each other. The encoding of lyrics and tunes together, therefore, generates emerging song properties, which then influence the upcoming predictions for both lyrics and tunes. This model applies to all three conditions we tested: music-only, speech-only, and song, as either the right or left branches of the model will be inactivated in unimodal scenarios. The model is potentially relevant to other types of "multimodal" scenarios and stimuli, such as poetry (combining speech and rhythm/rhyme) or watching dance (combining visual and musical signals). In the following paragraphs, we elaborate on how those common resources could be allocated to either the processing of linguistic or melodic information in song based on our findings.

### **Internal and external factors determining resource allocation**

We observed a trade-off in the neural encoding of speech and music expectations, revealed by an inverse correlation that emerged only in responses to song. This suggests that music and speech processing compete for the allocation of shared resources, in line with the idea that music and speech share processing regions [1]. Shared resources can be flexibly allocated to music or linguistic processing depending on the "internal" preferences of the listener and on stimulus properties that are often "external" to the listener's competence. Thus, during song listening, individuals with strong melodic encoding exhibited weaker phonemic encoding (internal factor). Similarly, songs eliciting a stronger melodic encoding often elicited weaker phonemic expectations (external factor). In the context of song, "internal" factors are determined by listeners' biases towards melodic or linguistic content reflecting their musical taste, culture, training, emotional state, and other factors impacting attentional resources. "External" biases, by contrast, reflect the balance of lyrics-tunes salience as intended by the composer or by the singer. Such a salience is not necessarily associated with an obvious acoustic predominance but is rather due to semantic or structural elements, such as melodic complexity. Music enjoyment, for example, is related to melodic complexity in an inverted U-shaped manner [40, 13] with entropy and surprisal predicting pleasure [37] and reward [38]. Such a relationship indicates that prediction mechanisms have a role to play beyond simple error minimization, e.g., explaining humans' curiosity and novelty-seeking nature [64].

We found that the trade-off between the encoding of phonemic and melodic expectations is inversely related to the listeners' musical sophistication measures. This may reflect limitations of the shared cognitive resources between speech and music processes, e.g. when allocated based on attention. Our task did not include any specific attentional requirements as participants could freely attend to the stimulus as a whole or to the parts they find most attractive, which allowed us to disentangle the internal (subjective) from the external (stimulus-related) factors that drive attention. Indeed, it would be interesting to extend this work to selective-attention tasks, where participants are asked to focus on either the speech or musical context while listening to songs. Such a study, with a sufficiently large and diverse participant cohort, could shed light on whether musical training enhances the interest in music structure or if there is also an increased pool of cognitive resources, enabling an enhanced simultaneous encoding of speech and music.

We also noted a positive correlation between the encoding of melodic expectations and "emotions factor" metrics that capture emotional responses to music [36]. Future research could explore these "internal" and "external" factors by including participants with known biases to speech or music (e.g., non-native speakers) or by exploring a larger variety of musical pieces with a wider range of salience between speech and musical components by manipulating the melodic and rhythmical complexity of the songs [13, 39].

### **Right hemispheric bias for melodic expectations in song processing**

A right hemispheric bias for melodic expectations was observed in response to song, which suggested that the brain employs distinct strategies when simultaneously processing tunes and lyrics. Both music and linguistic networks supporting predictive processing may extend bilaterally, but when handling parallel information, they may undergo re-weighting, leading to stronger spatial segregation of the two in the measured EEG responses.

Several studies have noted hemispheric asymmetries for speech and music [9, 65]. Yet, these biases only emerged when acoustical cues (e.g., temporal or spectral information) were degraded, thus impairing the processing of higher-level properties that are extracted from the acoustics in a bottom-up process. Here, we measured a similar asymmetry without manipulations, finding lateralization for high-level processes but not for acoustic properties. This is compatible with the previous finding that music, speech, and song neural selectivity is not explainable by acoustic features [10, 11].

Other investigations using the speech-to-song illusion suggested that hemispheric specialization is modulated by top-down factors that induce the listener to perceive the same vocalizations as song or speech even when acoustics properties are identical [52, 66, 67]. Such an effect was shown in the right mid-posterior STS/MTG, indicating a potential role in the integration of music and speech [52]. Such regions were already associated with song perception [61, 68], imagination [69], melody recognition [70], and classification of music from speech [71], indicating its role in both multimodal and unimodal processes. Altogether, these observations make STS/MTG a candidate for the right hemispheric bias we observe in the responses to songs.

### **Song features beyond phonemes and melody**

The results of this study have implications for current theories on music and speech cognition [17, 51, 14], paving the way for further investigation of the neural mechanisms specific to lyrics-tunes integration behind simple acoustic features (e.g., temporal vs. spectral). The melodic and phonemic expectation processes studied here only partly capture song processing. Future efforts will explore the contributions of other fundamental properties, such as rhythm, harmony, semantics, and syntax. For instance, while our choice of comparing melodic and phonemic expectations was motivated by their role as fundamental units operating on a similar time scale (acoustically invariant, just after

acoustics processing), one may consider other types of speech expectations like prosody that might be closer to melodic expectations in the sense that they imply an emotional and engagement element. Similarly, one may contrast melodic and semantic expectations to see if signatures of the internal and external drivers we found in our data become clearer, as it seems reasonable that both the internal preference for lyrics as well as the weight that the composer gave to them can be driven more by the semantics than by phonemes. This calls for further investigations with explicit within-subject comparisons, with the present findings and data providing a starting point to tackle these questions.

## Conclusion

This study probed the neural processing of speech and music expectations, providing evidence for their similarities, distinctions, and interactions. Our results complement the literature, supporting the view that speech and music expectations are induced in parallel by distinct yet interacting processes. The presented experimental framework allows for the investigation of multiple domains and their interaction, substantially contributing to the search for shared neural mechanisms at the foundation of lyrics-tunes integration. The model we propose could potentially apply to integration mechanisms involving other expressions (e.g., dance and music), sensory (e.g., audio-visual), or sensory-motor modalities (e.g., contrasting miming, speaking, and listening), as well as auditory sources (e.g., polyphonic music) into a coherent percept thus providing a framework for hypothesis formulation and testing.

## Materials and Method

### Participants

The study was conducted in accordance with the Declaration of Helsinki and was approved by the CNRIPH committee. Twenty healthy individuals with no history of hearing impairment or neurological disorders participated in the study (10 females, aged 23 to 51, mean age 28, with three left-handed individuals and one ambidextrous). All except one (who had a C1 English proficiency) were native English speakers with different degrees of musical training (7 participants had no musical background). Written informed consent was obtained from all participants, who were compensated for their participation. Each participant was tested in a single session and completed a general demographic test and the *Goldsmiths Musical Sophistication test* (Gold-MSI) to assess individual differences in musical listening abilities [36]. The Gold-MSI allows the assessment of musical sophistication independent of musical preferences for specific styles and, most importantly, is sensitive to individual differences among non-musicians. Furthermore, it can be factorized into facets of music sophistication: Active Engagement, Perceptual Abilities, Musical Training, Singing Abilities, and Emotions factors [36].

### Stimuli creation

The set of stimuli was explicitly built to have the most direct possible comparison among conditions and minimize possible confounding factors [34, 51]. This factor pertains to the differences in response strengths and patterns we might encounter when comparing linguistic and melodic components, which could arise from lower-level differences in the acoustics of speech and music (e.g., polyphonic music versus human voice) and not from different linguistic and non-linguistic demands. Hence, comparing speech with singing and humming produced by the same singer was the most natural choice. Specifically, we used the *NHSS Speech and Singing Parallel Database* [72], which collects a set of lyrics available in the sung and spoken version recorded by the same singer at  $44.1kHz$ , along with the phoneme-level alignment. To ensure a fair comparison, pairs of conditions had to include the same amount of information content (either melodic or linguistic). Therefore, we selected 18 lyrics with a minimum of 20 minutes of spoken content, corresponding to roughly 43 minutes of sung speech and the same amount for melodies. The corresponding melody was extracted from the sung speech waveform using pYIN [73], a modification of the YIN algorithm for fundamental frequency ( $F_0$ ) estimation [74]. In the first step of pYIN,  $F_0$  candidates and their probabilities are computed using the YIN algorithm. In the second step, Viterbi decoding is used to estimate the most likely  $F_0$  sequence and voicing flags. We limited the possible  $F_0$  candidates in the range  $C2(\sim 65Hz)$ - $C7(\sim 2093Hz)$  and set the frame and hop size for the *STFT* used for the pitch extraction to respectively 2048 and 128 samples at  $44.1kHz$ . The aim was to create a “hummed” version of the melody without any specific instrument timbre, especially those with sharp attacks like the piano. Previous research has shown that the brain is highly sensitive to percussive note onsets (characterized by an energy burst in the spectrogram), and we wanted to avoid this possible confound. To this end, we synthesized it using the extracted  $F_0$  and two harmonics, with a fade in/out and frequency interpolation transition length of  $10ms$  using *MeloSynth* [75]. Finally, to better mimic the human voice fluctuations and minimize acoustic confounds, we applied the energy envelope of the original signal and smoothed sharp attacks using a median filtering on the spectrogram [76, 77]. Finally, the musical scores were obtained using *Tony*, a note segmentation algorithm specific to singing voice

transcription that converts a pitch sequence into a set of discrete and quantized note events starting from the pYIN output [78]. Each note is then defined by its onset time, duration, and MIDI pitch value and saved into MIDI files. Thanks to Tony, a user interface aimed at streamlining the melody annotation process made available by the same authors [78], we could visually check and manually correct the annotations.

## Experimental procedure

Each participant listened to the same set of stimuli but in a different randomized order. To prevent participants from listening to different versions (melody, speech, and song) of the same song sequentially, we divided the trials into three main experimental blocks, where each subject listened to at least one version of each song. We included two main pauses between experimental blocks, during which the participants could take a break for 10 – 15 minutes. However, they were allowed to take short breaks between trials if necessary. To assess engagement with the stimuli, we included the following behavioral task: at the end of 20% of the trials, participants were asked a question about the stimulus they had just heard. For speech trials, the question was about the content, while for music trials, it was whether a little melody excerpt had been extracted from the stimulus they had just heard. The question could be one of two types for sung speech, with a 50% chance of each.

## Data acquisition

Participants listened to audio samples while sitting in a sound-proof, electrically shielded booth in dim light conditions. Audio stimuli were presented monophonically at a sampling rate of  $44.1kHz$  using Sennheiser HD650 headphones and Psychopy [79] customized Python code. Participants were instructed to maintain visual fixation on a crosshair centered on a screen and minimize motor activities. 64-channel EEG data and two extra electrodes on the mastoid bones were recorded and digitized at  $2048Hz$  using a BioSemi Active Two system. A customized analog system ensured optimal synchronization between the stimulus and the EEG responses.

## EEG preprocessing

EEG data were analyzed offline using MNE Python [80, 81] and Matlab (Mathworks). The preprocessing pipeline follows guidelines for linearly modeling neurophysiological data to continuous auditory stimuli [28]. First, the EEG was segmented into trials at the original sampling rate to avoid synchronization issues. Secondly, each trial was filtered between 1 and  $8Hz$  using low- and high-pass Butterworth zero-phase filters (order 3 with a forward and backward pass) and downsampled to  $64Hz$ . Channels with a variance exceeding three times that of the surrounding ones were replaced by an estimate calculated using spherical spline interpolation. All channels were re-referenced to the average of the two mastoid channels to maximize auditory responses [28]. The first  $500ms$  at the start of each trial were removed to avoid responses elicited by the start of the stimulus. Finally, the EEG responses of each participant were standardized together to preserve the relative power across channels.

## Predictor variables

In the present study, we aimed to quantify how much variance in the brain responses is explained by various features of the auditory stimuli, ranging from low-level acoustic features to higher-level probabilistic features reflecting expectations of fundamental speech and/or music units. The choice of which features to include among the predictor variables was made considering research work focused on either music [17, 18, 19] or speech perception [20, 21, 26]. All features were extracted from the waveforms and ground truth annotations that come along with each version of the stimulus (melody, song, or speech) and were downsampled at  $64Hz$  to match them precisely to the EEG signals.

Since stimuli are exclusively monophonic, the sound acoustic  $A$  was characterized by the broadband envelope  $E$ , computed as the absolute value of the waveform’s Hilbert transforms. Additionally, we included its halfway rectified first derivative  $D$  to model acoustic changes. Indeed, it is well known that brain responses are particularly sensitive to acoustic contrasts and bursts of energy. Phonemic information is often specifically located in acoustic onsets [82]; thus, it is important to control for responses to acoustic onsets [60, 20]. Similar considerations were made for music onsets [17]. On top of that, we included the onsets of the expectation features under test represented as one-hot encoded vectors [20, 17]. Phoneme onsets  $Op$  were included when testing hypotheses on phonemic expectation encoding, and note onsets  $Om$  when testing hypotheses on melodic expectation encoding. Phoneme onsets were determined from the phoneme-level annotations of the lyrics provided by the NHSS dataset, while note onsets were determined from the MIDI of musical score obtained using an automatic transcription algorithm [78], whose output was manually inspected and corrected (for more details see the section on stimuli creation). All expectation-related predictors were finally modeled as non-normalized modulations of these onsets.

### Phonemic expectation features

Phonemic expectation features, namely phonemic surprisal  $S_p$  and cohort entropy  $H_p$ , were computed for all phonemes  $p_k$  of a word of length  $K$  except for the starting one  $p_1$ , which was not modeled. First, the likelihood of a phoneme sequence  $p_1 \dots p_K$  being a valid speech token was estimated using a statistical model learned from a pronouncing dictionary. Given the *CMU Pronouncing Dictionary* [32], for each subsequence of phonemes, or cohort,  $p_1 \dots p_k$  with  $1 < k \leq K$ , we computed the probability distribution of the next possible phoneme  $p_{k+1}$  by counting compatible occurrences in the dictionary. Given the probability distribution, the surprisal of phoneme  $k$  describes how phoneme  $p_k$  is surprising/unexpected given the previous phoneme sequence  $p_1 \dots p_{k-1}$  in the word, or, in other words, it expresses its information content. Specifically, it is computed as the logarithm of the inverse conditional probability of occurrence of a given phoneme  $p_k$  after a given phoneme sequence  $p_1 \dots p_{k-1}$ :

$$S(p_k | p_1 \dots p_{k-1}) = \log_2 \left( \frac{1}{P(p_k | p_1 \dots p_{k-1})} \right)$$

The second feature that was extrapolated is the entropy of a given phonemic cohort/context. This measure was defined as the Shannon entropy [33] and computed by averaging the surprise over all possible phonemes in an alphabet  $\mathcal{E}$  that can serve as continuations of that phonemic sequence:

$$H(p_1 \dots p_k) = \sum_{p_{k+1} \in \mathcal{E}} P(p_{k+1} | p_1 \dots p_k) S(p_{k+1} | p_1 \dots p_k)$$

These two features are complementary in that the first indicates the degree of expectedness, while the second clarifies the degree of uncertainty of the context, thus corresponding to a weaker or stronger sequence violation, respectively. We did not use the same definition of phoneme surprisal and cohort entropy as given by Brodbeck in [20], as we wanted to be consistent with the definitions used for music sequences given by Di Liberto in [17], which are based on the concepts of predictability and uncertainty and their role in musical pleasure and learning [37, 83, 48].

### Melodic expectation features

Therefore, we defined the melodic surprisal  $S_m$  and entropy  $H_m$  exactly as we did for phonemes. Given a note  $n_k$ , a note sequence preceding that note  $n_1 \dots n_{k-1}$ , and an alphabet  $\mathcal{E}$  describing all possible pitch values,  $S(n_k | n_1 \dots n_{k-1})$  refers to the inverse probability of occurrence of a particular note at a given position in the melody and  $H(n_k)$  describes how much that note sequence or context is uncertain. To estimate the probability distributions of each upcoming note  $n_k$  necessary for the computation of  $S_k$  and  $H_k$ , we used *Information dynamics of music* (IDyOM), a statistical model for implicit melodic learning based on variable-order Markov chains [31], whose modeling of human melodic expectation is already supported by several behavioral and neural studies [17, 31, 84, 85, 86]. This model can operate on multiple viewpoints, meaning that it can capture the distributions of various properties of music. Here, we focused on the pitch of a note and note on the onset time as done by Di Liberto et al. [17], as we wanted a parallel comparison of phonemic and melodic expectation encoding. IDyOM causally predicts upcoming notes in a melodic sequence by weighting predictions from a long-term memory (LTM) statistical model learned from a large corpus of Western music and short-term memory statistics (STM) learned online on the current musical piece [31]. While the first one simulates the statistical knowledge of a listener implicitly acquired in a lifetime of exposure to music, the second one is associated with the short-term memory of the statistics of the musical piece under test [31]. Specifically, the STM model estimates the probability distribution over every possible note continuation for every  $n$ -gram context up to a given length  $k$  (model order), which in our case is set to 20, and combined with an entropy-based weighting function [31]. We kept the model order low as opposed to Di Liberto and colleagues [17], as we wanted to capture melodic surprisal only related to short contexts and avoid capturing musical structure at multiple hierarchical levels (defined as context lengths).

### Model estimation

For each participant and condition, the channel-specific mapping between a given feature set and the neural data was estimated by solving a regularised linear regression problem [27]. Therefore, separate and independent optimal filters were estimated for each of the 64 channels, 20 participants, and three conditions while considering all predictor variables concurrently. Nested leave-one-out cross-validation (across trials) was used to assess how well the models could predict unseen data while controlling for overfitting and selecting the regularisation parameter lambda. The model is learned considering multiple stimulus-response time lags tau to account for non-instantaneous interactions [27]. We

first fit the model using an exploratory time-lag window ranging from  $-200$  to  $1000ms$ , which was later restricted to the significant peaks between  $-100$  and  $700ms$  for the final analysis, thus reducing the dimensionality of the data and the risk of overfitting. Note that when training, the considered time window is always  $[\tau_{min} - \delta, \tau_{max} + \delta]$ , where  $\delta = 50ms$  to avoid border artifacts. The quality of a prediction was then quantified by calculating the z-transformed Pearson’s correlation between the measured signals and the corresponding predictions at each scalp electrode. The Fisher z-transformation is the inverse hyperbolic tangent of Pearson’s correlation coefficient and is applied to correct for distortions introduced by the fixed end-points at 1 and  $-1$ , yielding a variable whose distribution is approximately normally distributed with stable variance optimal for hypothesis testing. Model estimation and cross-validation algorithms are publicly available in the `mTRF toolbox` open-source Matlab package [27].

## Models comparison

We hypothesized that cortical signals encode melodic or phonemic expectations or both depending on the condition under test (speech, melody, and song). Consequently, we expected more accurate EEG predictions when using a combined feature-set  $AOpP$  for speech and song and  $AOmM$  for melody and song, as opposed to relying solely on acoustic features  $A$  and onsets  $O$ . The reasoning behind this is that including entropy and surprisal predictors should improve the fitting score if the EEG responses are modulated by expectations, that is if  $P$  and  $M$  describe EEG dynamics that are not redundant with  $AO$ . To assess whether a given predictor brings significant contributions, we always compared the predictive performance of the complete model to the performance of a baseline model that adheres to the null hypothesis. This baseline is identical to the complete model except for the predictor under investigation, which is randomly shuffled in a manner that is appropriate to the hypothesis being tested. A predictor was considered to be encoded in the EEG if it significantly improved the model’s fit across participants. Under the null hypothesis, which posits no causal relationship between the given predictor and the responses, a shuffled version of the predictor should perform equally well as the properly aligned version. Consequently, any difference in model fit between the full model and the shuffled models thus indicates a significant relationship between predictor and responses. This procedure enabled us to evaluate any incremental improvement in the model’s performance due to a particular predictor without introducing bias by altering the model’s complexity (the size of the predictors) and has been widely employed in cortical tracking experiments [17].

## Test of lateralisation

Tests of hemispheric asymmetry were performed by comparing model-fit improvement between the two hemispheres electrode by electrode, similarly as in [20]. First, a difference map was computed by subtracting from the z-score values of the full model the z-score values of the shuffled model. The left and right hemispheres were then compared electrode-wise by mapping the electrodes of one hemisphere into the other in a mirrored fashion and computing a two-tailed t-test while controlling for multiple comparisons. A potential confounding factor could be biases due to noise or other unknown factors toward one of the two hemispheres. However, a lateralization test of the reduced model (only acoustics) indicated no overall difference in predictive power in the left and right hemispheres.

## Statistical analysis

Statistical analyses were performed using two-tailed t-tests or non-parametric Wilcoxon signed-rank tests for pair-wise comparisons. The choice of using either the parametric or non-parametric test was based on the normality of the data, which was assessed via the Anderson-Darling test. The measure of the effect size was chosen similarly: Cohen’s  $d$  was used when the assumption of normality holds, Wilcoxon’s  $Z$  score otherwise. Correction for multiple comparisons was applied where necessary via the false discovery rate (FDR) approach. One-way ANOVA was used to assess when testing the significance of an effect over multiple ( $> 2$ ) groups. The statistical significance of TRFs was assessed with a one-sample t-test against 0 for every time lag. The subject and song-wise statistical analyses were done using a jackknife cross-validation procedure to control for outliers.

## Acknowledgments

This project was supported by an Advanced European Research Council grant (NEUME, 787836) and FrontCog grant ANR-17-EURE-0017. G.D.L. was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106\_P2 at the ADAPT SFI Research Centre at Trinity College Dublin. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme. The research in this paper used the NHSS Database made available by the HLT lab at the National University of Singapore. We thank G. Varni, R. McWalter, and A. Prosper for participating in the pilots.



## References

- [1] Aniruddh D Patel. Language, music, syntax and the brain. *Nature neuroscience*, 6(7):674–681, 2003.
- [2] Charles Darwin. The descent of man, and selection in relation to sex, 1871.
- [3] S Mithen. The singing neanderthals: The origins of music, language, mind, and body. *Harvard Univ Press, Cambridge, MA*, page 374, 2006.
- [4] Nobuo Masataka. The origins of language and the evolution of music: A comparative perspective. *Physics of Life Reviews*, 6:11–22, 3 2009.
- [5] Liberty S Hamilton. Human song: Separate neural pathways for melody and speech. *Current Biology*, 32(7):R311–R313, 2022.
- [6] Ilana Harris, Efe C. Niven, Alex Griffin, and Sophie K. Scott. Is song processing distinct and special in the auditory cortex? *Nature Reviews Neuroscience* 2023 24:11, 24:711–722, 10 2023.
- [7] Daniela Sammler, Ameer Baird, Romain Valabrègue, Sylvain Clément, Sophie Dupont, Pascal Belin, and Séverine Samson. The relationship of lyrics and tunes in the processing of unfamiliar songs: a functional magnetic resonance adaptation study. *Journal of Neuroscience*, 30(10):3572–3578, 2010.
- [8] Corianne Rogalsky, Feng Rong, Kourosh Saberi, and Gregory Hickok. Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. *Journal of Neuroscience*, 31(10):3843–3852, 2011.
- [9] Robert J. Zatorre, Pascal Belin, and Virginia B. Penhune. Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, 6:37–46, 1 2002.
- [10] Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *neuron*, 88(6):1281–1296, 2015.
- [11] Sam V Norman-Haignere, Jenelle Feather, Dana Boebinger, Peter Brunner, Anthony Ritaccio, Josh H McDermott, Gerwin Schalk, and Nancy Kanwisher. A neural population selective for song in human auditory cortex. *Current Biology*, 32(7):1470–1484, 2022.
- [12] Micha Heilbron and Maria Chait. Great expectations: is there evidence for predictive coding in auditory cortex? *Neuroscience*, 389:54–73, 2018.
- [13] Peter Vuust, Ole A. Heggli, Karl J. Friston, and Morten L. Kringelbach. Music in the brain. *Nature Reviews Neuroscience* 2022 23:5, 23:287–305, 3 2022.
- [14] Stefan Koelsch, Peter Vuust, and Karl Friston. Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, 23:63–77, 1 2019.
- [15] Allison R Fogel, Jason C Rosenberg, Frank M Lehman, Gina R Kuperberg, and Aniruddh D Patel. Studying musical and linguistic prediction in comparable ways: The melodic cloze probability method. *Frontiers in psychology*, 6:154672, 2015.
- [16] Aniruddh D Patel and Emily Morgan. Exploring cognitive relations between prediction in language and music. *Cognitive Science*, 41:303–320, 2017.
- [17] Giovanni M Di Liberto, Claire Pelofi, Roberta Bianco, Prachi Patel, Ashesh D Mehta, Jose L Herrero, Alain de Cheveigné, Shihab Shamma, and Nima Mesgarani. Cortical encoding of melodic expectations in human temporal cortex. *Elife*, 9:e51784, 2020.
- [18] Guilhem Marion, Giovanni M Di Liberto, and Shihab A Shamma. The music of silence: part i: responses to musical imagery encode melodic expectations and acoustics. *Journal of Neuroscience*, 41(35):7435–7448, 2021.
- [19] Giovanni M Di Liberto, Guilhem Marion, and Shihab A Shamma. The music of silence: Part ii: Music listening induces imagery responses. *Journal of Neuroscience*, 41(35):7449–7460, 2021.
- [20] Christian Brodbeck, L Elliot Hong, and Jonathan Z Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018.
- [21] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P De Lange. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119, 2022.
- [22] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12, 2023.
- [23] Giovanni M Di Liberto, Daniel Wong, Gerda Ana Melnik, and Alain de Cheveigné. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. *Neuroimage*, 196:237–247, 2019.

- [24] Giovanni M Di Liberto, Jingping Nie, Jeremy Yeaton, Bahar Khalighinejad, Shihab A Shamma, and Nima Mesgarani. Neural representation of linguistic feature hierarchy reflects second-language proficiency. *Neuroimage*, 227:117586, 2021.
- [25] Laura Gwilliams, Jean-Remi King, Alec Marantz, and David Poeppel. Neural dynamics of phoneme sequences reveal position-invariant code for content and order. *Nature communications*, 13(1):6606, 2022.
- [26] Marlies Gillis, Jonas Vanthornhout, Jonathan Z Simon, Tom Francart, and Christian Brodbeck. Neural markers of speech comprehension: measuring eeg tracking of linguistic speech representations, controlling the speech acoustics. *Journal of Neuroscience*, 41(50):10316–10329, 2021.
- [27] Michael J Crosse, Giovanni M Di Liberto, Adam Bednar, and Edmund C Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10:604, 2016.
- [28] Michael J Crosse, Nathaniel J Zuk, Giovanni M Di Liberto, Aaron R Nidiffer, Sophie Molholm, and Edmund C Lalor. Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research. *Frontiers in Neuroscience*, page 1350, 2021.
- [29] Liberty S Hamilton and Alexander G Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience*, 35(5):573–582, 2020.
- [30] Giovanni M Di Liberto, James A O’sullivan, and Edmund C Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.
- [31] Marcus Thomas Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition*. PhD thesis, City University London, 2005.
- [32] Robert L Weide. The cmu pronouncing dictionary. URL: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>, 1998.
- [33] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
- [34] Giorgia Cantisani, Amirhossein Chalehchaleh, Giovanni Di Liberto, and Shihab Shamma. Investigating the cortical tracking of speech and music with sung speech. In *Proc. INTERSPEECH 2023*, pages 5157–5161. ISCA, 2023.
- [35] Xiaoyu Tang, Jinglong Wu, and Yong Shen. The interactions of multisensory integration with endogenous and exogenous attention. *Neuroscience & Biobehavioral Reviews*, 61:208–224, 2016.
- [36] D Müllensiefen, B Gingras, J Musil, and L Stewart. Measuring the facets of musicality: the goldsmiths musical sophistication index (gold-msi). *Personality and Individual Differences*, 60:S35, 2014.
- [37] Vincent KM Cheung, Peter MC Harrison, Lars Meyer, Marcus T Pearce, John-Dylan Haynes, and Stefan Koelsch. Uncertainty and surprise jointly predict musical pleasure and amygdala, hippocampus, and auditory cortex activity. *Current Biology*, 29(23):4084–4092, 2019.
- [38] Valorie N Salimpoor, David H Zald, Robert J Zatorre, Alain Dagher, and Anthony Randal McIntosh. Predictions and the brain: How musical sounds become rewarding. *Trends in cognitive sciences*, 19(2):86–91, 2015.
- [39] Olivier Senn. A predictive coding approach to modelling the perceived complexity of popular music drum patterns. *Heliyon*, 9(4), 2023.
- [40] Guy Madison and Gunilla Schiölde. Repeated listening increases the liking for music regardless of its complexity: Implications for the appreciation and aesthetics of music. *Frontiers in neuroscience*, 11:235557, 2017.
- [41] Matthew K Leonard, Maxime O Baud, Matthias J Sjerps, and Edward F Chang. Perceptual restoration of masked speech in human cortex. *Nature communications*, 7(1):13619, 2016.
- [42] Dennis Norris, James M McQueen, and Anne Cutler. Prediction, bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience*, 31(1):4–18, 2016.
- [43] Yaara Yeshurun, Stephen Swanson, Erez Simony, Janice Chen, Christina Lazaridi, Christopher J Honey, and Uri Hasson. Same story, different story: the neural representation of interpretive frameworks. *Psychological science*, 28(3):307–319, 2017.
- [44] Carolyn McGettigan, Andrew Faulkner, Irene Altarelli, Jonas Obleser, Harriet Baverstock, and Sophie K Scott. Speech comprehension aided by multiple modalities: behavioural and neural interactions. *Neuropsychologia*, 50(5):762–776, 2012.
- [45] Antje Strauß, Sonja A Kotz, and Jonas Obleser. Narrowed expectancies under degraded speech: revisiting the n400. *Journal of Cognitive Neuroscience*, 25(8):1383–1395, 2013.

- [46] Robert J Zatorre and Valorie N Salimpoor. From perception to pleasure: music and its neural substrates. *Proceedings of the National Academy of Sciences*, 110(supplement\_2):10430–10437, 2013.
- [47] Elizabeth Hellmuth Margulis. *On repeat: How music plays the mind*. Oxford University Press, USA, 2014.
- [48] Benjamin P Gold, Marcus T Pearce, Ernest Mas-Herrero, Alain Dagher, and Robert J Zatorre. Predictability and uncertainty in the pleasure of music: a reward for learning? *Journal of Neuroscience*, 39(47):9397–9409, 2019.
- [49] W Tecumseh Fitch and Mauricio D Martins. Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, 1316(1):87–104, 2014.
- [50] Christopher C Heffner and L Robert Slevc. Prosodic structure as a parallel to musical structure. *Frontiers in psychology*, 6:1962, 2015.
- [51] Isabelle Peretz, Dominique Vuvan, Marie-Élaine Lagrois, and Jorge L Armony. Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1664):20140090, 2015.
- [52] Mark Hymers, Garreth Prendergast, Can Liu, Anja Schulze, Michellie L Young, Stephen J Wastling, Gareth J Barker, and Rebecca E Millman. Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, 108:225–233, 2015.
- [53] Nathaniel J Zuk, Jeremy W Murphy, Richard B Reilly, and Edmund C Lalor. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLoS computational biology*, 17(9):e1009358, 2021.
- [54] Stefan Koelsch. Music-syntactic processing and auditory memory: Similarities and differences between eran and mmm. *Psychophysiology*, 46(1):179–190, 2009.
- [55] Peter Vuust, Elvira Brattico, Miia Seppänen, Risto Näätänen, and Mari Tervaniemi. The sound of music: differentiating musicians using a fast, musical multi-feature mismatch negativity paradigm. *Neuropsychologia*, 50(7):1432–1443, 2012.
- [56] Mireille Besson and Françoise Macar. An event-related potential analysis of incongruity in music and other non-linguistic contexts. *Psychophysiology*, 24(1):14–25, 1987.
- [57] Ghislaine Dehaene-Lambertz, Emmanuel Dupoux, and Ariel Gout. Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of cognitive neuroscience*, 12(4):635–647, 2000.
- [58] John F Connolly and Natalie A Phillips. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of cognitive neuroscience*, 6(3):256–266, 1994.
- [59] Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.
- [60] Christoph Daube, Robin AA Ince, and Joachim Gross. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, 29(12):1924–1937, 2019.
- [61] Daniele Schön, Reyna Leigh Gordon, and Mireille Besson. Musical and linguistic processing in song perception. *Annals of the New York Academy of Sciences*, 1060(1):71–81, 2005.
- [62] Julia Merrill, Daniela Sammler, Marc Bangert, Dirk Goldhahn, Gabriele Lohmann, Robert Turner, and Angela D Friederici. Perception of words and pitch patterns in song and speech. *Frontiers in psychology*, 3:18675, 2012.
- [63] Ilana Harris, Efe C Niven, Alex Griffin, and Sophie K Scott. Is song processing distinct and special in the auditory cortex? *Nature Reviews Neuroscience*, 24(11):711–722, 2023.
- [64] Andy Clark. A nice surprise? predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, 17(3):521–534, 2018.
- [65] Philippe Albouy, Lucas Benjamin, Benjamin Morillon, and Robert J Zatorre. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, 367(6481):1043–1047, 2020.
- [66] Felix Haiduk, Robert J Zatorre, Lucas Benjamin, Benjamin Morillon, and Philippe Albouy. Spectrotemporal cues and attention jointly modulate fmri network topology for sentence and melody perception. *Scientific Reports*, 14(1):5501, 2024.
- [67] Adam Tierney, Fred Dick, Diana Deutsch, and Marty Sereno. Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, 23(2):249–254, 2013.
- [68] Daniele Schön, Reyna Gordon, Aurélie Campagne, Cyrille Magne, Corine Astésano, Jean-Luc Anton, and Mireille Besson. Similar cerebral networks in language, music and song perception. *Neuroimage*, 51(1):450–461, 2010.

- [69] Robert J Zatorre and Andrea R Halpern. Effect of unilateral temporal-lobe excision on perception and imagery of songs. *Neuropsychologia*, 31(3):221–232, 1993.
- [70] Isabelle Peretz, Nathalie Gosselin, Pascal Belin, Robert J Zatorre, Jane Plailly, and Barbara Tillmann. Music lexical networks: the cortical organization of music recognition. *Annals of the New York Academy of Sciences*, 1169(1):256–265, 2009.
- [71] Daniel A Abrams, Anjali Bhatara, Srikanth Ryali, Evan Balaban, Daniel J Levitin, and Vinod Menon. Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. *Cerebral Cortex*, 21(7):1507–1518, 2011.
- [72] Bidisha Sharma, Xiaoxue Gao, Karthika Vijayan, Xiaohai Tian, and Haizhou Li. Nhss: A speech and singing parallel database. *Speech Communication*, 133:9–22, 2021.
- [73] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *2014 IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*, pages 659–663. IEEE, 2014.
- [74] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.
- [75] Justin Salomon. Melosynth, 2018.
- [76] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx)*, volume 13, pages 1–4, 2010.
- [77] Jonathan Driedger, Meinard Müller, and Sascha Disch. Extending harmonic-percussive separation of audio signals. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 611–616, 2014.
- [78] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, and Simon Dixon. Computer-aided melody note transcription using the tony software: Accuracy and efficiency, 2015.
- [79] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51:195–203, 2019.
- [80] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, page 267, 2013.
- [81] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.
- [82] Keith R Kluender, Jeffrey A Coady, and Michael Kiefte. Sensitivity to change in perception of speech. *Speech communication*, 41(1):59–69, 2003.
- [83] Marcus T Pearce, María Herrojo Ruiz, Selina Kapasi, Geraint A Wiggins, and Joydeep Bhattacharya. Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, 50(1):302–313, 2010.
- [84] Roberta Bianco, Lena Esther Ptasczynski, and Diana Omigie. Pupil responses to pitch deviants reflect predictability of melodic sequences. *Brain and Cognition*, 138:103621, 2020.
- [85] Niels Chr Hansen and Marcus T Pearce. Predictive uncertainty in auditory sequence processing. *Frontiers in psychology*, 5:1052, 2014.
- [86] Toviah Moldwin, Odelia Schwartz, and Elyse S Sussman. Statistical learning of melodic patterns influences the brain’s response to wrong notes. *Journal of cognitive neuroscience*, 29(12):2114–2122, 2017.